Neural Networks with Feature Attribution and Contrastive Explanations

Housam K. B. Babiker[1,3] (✉), Mi-Young Kim[2,3], and Randy Goebel[1,3]

[1] Department of Computing Science, University of Alberta
[2] Department of Science, Augustana Faculty, University of Alberta
[3] Alberta Machine Intelligence Institute
{khalifab,miyoung2,rgoebel}@ualberta.ca

Abstract. Interpretability is becoming an expected and even essential characteristic in GDPR Europe. In the majority of existing work on natural language processing (NLP), interpretability has focused on the problem of explanatory responses to questions like "*Why p?*" (identifying the causal attributes that support the prediction of "*p*.)" This type of local explainability focuses on explaining a single prediction made by a model for a single input, by quantifying the contribution of each feature to the predicted output class. Most of these methods are based on post-hoc approaches. In this paper, we propose a technique to learn centroid vectors concurrently while building the black-box in order to support answers to "*Why p?*" and "*Why p and not q?*," where "*q*" is another class that is contrastive to "*p*." Across multiple datasets, our approach achieves better results than traditional post-hoc methods.

Keywords: Interpretability · NLP · Text classification.

1 Introduction

Research on making deep learning models more interpretable and explainable is receiving much attention. One of the main reasons is the application of deep learning models to high-stake domains. In general, interpretability is an essential component for deploying deep learning models. Interpretability in the context of deep learning can be used to tackle a variety of problems: (i) the detection of biased views in a deep learning model, (ii) evaluation of the fairness of a deep learning model, (iii) faithfully explaining the predictions of the classifier, i.e., the construction of accurate explanation that explains the underlying causal phenomena [13] and (iv) the use of explanations as a proxy for model debugging, which allows researchers/engineers to construct models better or debug existing models. Non-linear deep neural networks come at the cost of model interpretability. Most existing related research has focused on identifying feature attribution (e.g., possible causal attributes) to explain the prediction of a black-box neural network. This type of explanation is defined as answers to "*why-questions.*" "*Why-questions*," are generally thought of as causal-like explanations [11]. Existing techniques to *why-questions* rely on using a post-hoc approach to identify the causal attributes for a single black-box prediction. Post-hoc methods generally do not always provide accurate explanations [20]. There are many possible reasons for this limitation; for instance, feature attributions typically suffer from noisy gradients in back-propagation techniques [8]. Studies in philosophy and social science show that humans, in general, prefer contrastive explanations, i.e., the explanation of an event is based on explaining the fact (p) in contrast to another event (q) [16, 12]. Here "p" represents the model prediction, and "q" represents an alternative class we would use for a contrastive explanation. A contrastive explanation is an essential property of an explanation: 1) humans ask a contrastive question when they are surprised by an event and expect a different outcome, and 2) the contrastive event is what they expect to happen [12, 16, 7]. Majority of existing post-hoc techniques are only limited to providing answers to "*why p?*" and cannot provide answers to "*why p and not q?*". For instance, gradient-based methods. Contrastive explanations are relatively new in NLP [9]. Our work focuses on building an inherently interpretable model that can support answers to both kinds of questions: "*why p?,*" and "*why p, not q?.*" In general, a contrastive explanation provides an explanation for why an instance had the current output (fact) rather than a targeted outcome of interest (foil) [25]. An example of our proposed neural network model is shown in Figure 1.

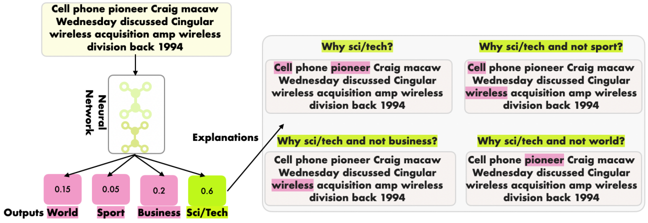
Cell phone pioneer Craig macaw Wednesday discussed Cingular wireless acquisition amp wireless division back 1994
Neural Network
Explanations
0.15
0.05
0.2
0.6
Outputs World
Sport
Business
Sci/Tech
Why sci/tech?
Cell phone pioneer Craig macaw Wednesday discussed Cingular wireless acquisition amp wireless division back 1994
Why sci/tech and not sport?
Cell phone pioneer Craig macaw Wednesday discussed Cingular wireless acquisition amp wireless division back 1994
Why sci/tech and not business?
Cell phone pioneer Craig macaw Wednesday discussed Cingular wireless acquisition amp wireless division back 1994
Why sci/tech and not world?
Cell phone pioneer Craig macaw Wednesday discussed Cingular wireless acquisition amp wireless division back 1994

Fig. 1. An example of the proposed neural network model with answers to "*why p?*" and "*why p and not q?*" questions. Here we visualize the top salient attributes.

1.1 Contrastive vs. Counterfactual

The evolving discussions of explainable AI (XAI) have articulated several distinguishing aspects of explanation (e.g., [16]), including a difference between contrastive (e.g., what made the *difference* between the students who failed the exam and those who did not fail?) and counterfactual (e.g., will we reduce climate change *if we reduce fuel consumption?*) explanations. Contrastive explanations are different from counterfactual explanations [15]. In general, contrastive and counterfactual reasoning emphasize different aspects of causation[5]. In counterfactual reasoning, we focus on instances in which the salient causal attributes are absent (missing from the text). In contrast, in a contrastive explanation (our focus here), one considers the difference in attributes between two predictions. The difference between the two approaches is in the knowledge support required for the explanation. For instance, a counterfactual explanation focuses on the question of "*What if?*," while a contrastive explanation focuses on "the difference."

The contributions of this paper can be summarized as follows: (i) we propose an interpretable (intrinsic) neural model that focuses on learning deep discriminative embedding features, (ii) our neural model provides two types of explanations (e.g., non-contrastive explanations and contrastive explanations) using feature attribution, and (iii) we proposed a metric to evaluate the quality of the contrastive explanations. An intrinsic neural model is better than using traditional post-hoc explanations because: (i) we can find faithful explanations, (ii), we do not need an additional complex computation to find an explanation for a single prediction.

2 Related work

2.1 Contrastive explanations

With contrastive explanations, we aim to expose an alternative to any given model prediction. In [9], they proposed a post-hoc approach that relies on a projection matrix to devise explanations. Similarly, [18] used SHAP to generate a contrastive explanation. Our approach is different; we propose an intrinsic neural model which supports answers to "*why p?*" and "*why p and not q?*" questions, rather than relying on post-hoc approaches. In the context of contrastive explanations, we focus on finding the difference in the attributes that could distinguish the prediction "*p*" from the foil "*q*."

2.2 Counterfactual explanations

Counterfactual explanations consist in generating text as a counterfactual example. In general, counterfactual explanations seek to identify a minimal change in model data that "flips" a predictive model's prediction, which is used for explanation. [26] proposed the concept of unconditional counterfactual explanations and introduced a framework for generating counterfactual explanations. For text classification, [29] proposed a method to generate counterfactual text from a pretrained model for the finance domain. In addition, [6] relied on finding evidence that is discriminative for the target class but not present in the foil class to learn a model to generate counterfactual explanations for why a model predicts class "*p*" instead of "*q*." However, their approach was mainly designed for computer vision.

2.3 Post-hoc non-contrastive explanations

One of the most popular techniques for explaining the prediction of a black box is the use of *why p?*. There is much prior work on this topic. For instance, [3] used Shapley approximation and proposed two methods, namely L-Shapley and C-Shapley. Additionally, [22] proposed the integrated gradient method, which relies on using a back-propagation algorithm. Other methods also rely on perturbation techniques such as [19]. Some methods focus on constructing interpretable neural architecture for classification. For instance, [2]'s model learn a rationale as the model's explanation. In general, our approach is different from traditional post-hoc and rationale-based models. We provide two types of explanations using an intrinsic neural model i.e., answers to "*why p?* and *why p and not q?*" questions. Overall, our work is not the first contribution to contrastive explanation nor the first technique for "*why p?*" questions. In [1], authors proposed a knowledge distillation technique which could learn an interpretable vector space model. However our work is different, we focus on building an intrinsic model which can support answers to why p and and *why p?*and *why p and not q?* questions.

3 Contrastive explanation generation

Our approach is not a post-hoc technique for model's explanation but rather the pursuit of constructing an inherently interpretable neural network. Our intrinsic neural model relies on improving the embedding features (see Figure 2.) For a given class in the dataset, our network attempts to assign similar texts into a single cluster.

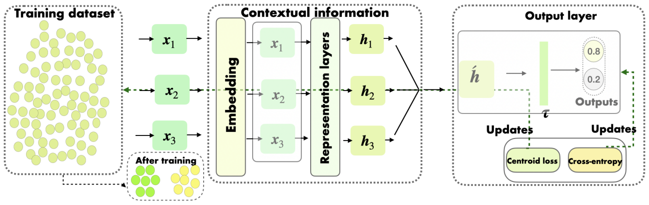

Fig. 2. Our proposed intrinsic neural architecture focuses on clustering texts based on the model predictions. For each class, we define a centroid vector. During training, we used our proposed method to establish a similarity structure between the sentences and the corresponding centroid vector.

3.1 Neural nets with feature attributions and contrastive explanations

We propose a multi-task neural network architecture, i.e., a classification task and an explanation task. We jointly optimize the network for both classifications and faithful explanations. For notation, we denote scalars with italic lowercase letters (e.g., x), vectors with bold lowercase letters (e.g., $\boldsymbol{x}$), and matrices with bold uppercase letters (e.g., $\boldsymbol{W}$). In the text classification task, an input sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_l \in \mathbb{R}^d$, where l is the length of the text input and d is the vector dimension, is mapped to a distribution over class labels using a parameterized neural network (e.g., a Multi-head attention). In general, the contextual vector $\acute{\boldsymbol{h}} \in \mathbb{R}^d$ is passed to a linear layer with parameters $\boldsymbol{W} \in \mathbb{R}^{d\times n}$ which provides a probability distribution over n classes. The output $\boldsymbol{y}$ is a vector of class probabilities of dimension $\mathbb{R}^n$, where n is the number of classes. The predicted label p of the text input is the index of the maximum element in $\boldsymbol{y}$, i.e., $p = argmax f(\boldsymbol{x})$, $\forall k \in [1, n]$. Here, k iterates over the probabilities and $f(\boldsymbol{x})$ denotes a neural network. During training, an empirical loss (e.g., cross-entropy) $\mathcal{J}(p, y^{'}, \theta)$ is minimized using gradient descent, where $y^{'}$ is the ground truth label and θ represents the network's parameters. We propose to augment the network to provide two types of explanations "*Why p?*" and "*Why p and not q?*." To do so, we first define a randomly initialized centroid vector for each class, and then use the centroid vector as a proxy to explain the black-box prediction.

For instance, if the neural network's prediction is class 1, we use the centroid vector representing that class to calculate the the scores for *why p?*. For contrastive explanation, we find the difference between the scores of the centroid vector representing the predicted class and the scores of the centroid vector representing the contrast class (e.g., the centroid vector for class 2). The centroid vector of label p pulls the weighted sentence vector of the text input $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_l$ closer. In the following, we discuss the steps for augmenting a neural network with the centroid vectors. Let $\boldsymbol{c}_j (j = 1, 2, ..., n)$ be a collection of randomly initialized centroid vectors, where $\boldsymbol{c}_j \in \mathbb{R}^d$ is a vector representing label $\boldsymbol{y}_j$. We propose a new objective function, namely centroid-loss, to explain the neural network predictions effectively. Our solution enhances the discriminative power of the deeply learned features in neural networks. Specifically, we learn a centroid $\boldsymbol{c}_j$ (a vector with the same dimension as an embedding feature) of each class. In the course of training, we simultaneously update the centroid vector and minimize the distances between the embedding features and their corresponding class' centroid vector.

3.2 Joint objective

Recall that a supervised learning algorithm input is a set of training instances and the corresponding label. The goal is to learn a function that accurately maps input examples to their desired labels using cross-entropy. Given the prediction p, we learn $\boldsymbol{c}_p \in \mathbb{R}^d$ to pull the sentence vectors representing class p closer. Intuitively, we are minimizing the intra-class variations while keeping the features of different classes separable. In the following, we discuss the optimization objective of our proposed network.

Cross-entropy: term 1 in the optimization objective function is the standard loss function for classification. We denote this loss as $\mathcal{L}^{\text{cls}}$.

Attractive term: term 2 focuses on minimizing the cosine distance between the sentence vector and the corresponding $\boldsymbol{c}_p$. Let $\boldsymbol{X}$ be a matrix consisting of embedding vectors $[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_l]$ and the sentence vector of $\boldsymbol{X}$ is $\hat{\boldsymbol{x}} \in \mathbb{R}^d$. Let, $\bar{\boldsymbol{w}} \in \mathbb{R}^l$ be the importance scores, where each.

$$\bar{\boldsymbol{w}}_i = \frac{\boldsymbol{x}_i \cdot \hat{\boldsymbol{x}}}{\|\boldsymbol{x}_i\| \, \|\hat{\boldsymbol{x}}_i\|}, \tag{1}$$

where $\bar{\boldsymbol{w}}_i$ is the importance score of word i, and $\hat{\boldsymbol{x}}$ is the sentence vector of the input $\boldsymbol{X}$. Term 2 minimizes the cosine distance between the weighted sentence vector $\bar{\boldsymbol{x}}$ of each input with the corresponding centroid vector $\boldsymbol{c}_p$. The sentence vector is defined as follows:

$$\bar{\boldsymbol{x}} = \boldsymbol{X}\left(\frac{\exp(\bar{\boldsymbol{w}})}{\sum_{i=1}^{l}\exp(\bar{\boldsymbol{w}}_i)}\right) \tag{2}$$

From equation 2, we obtain the weighted sentence vector through multiplying the values in the i-th row of $\boldsymbol{X}$ by $\bar{\boldsymbol{w}}_i$ followed by calculating the sentence vector $\bar{\boldsymbol{x}} \in \mathbb{R}^d$. We define the loss of term 2 as follows:

$$\mathcal{L}^{\mathrm{attr}} = 1 - \frac{\bar{\boldsymbol{x}} \cdot \boldsymbol{c}_p}{\|\bar{\boldsymbol{x}}\| \, \|\boldsymbol{c}_p\|} \tag{3}$$

Term 2 is the second loss of our proposed optimization objective.

Repulsive term: term 3 (the third term in the overall loss function) focuses on maximizing cosine distance of $\bar{\boldsymbol{x}}$ from other centroid vectors, i.e., $\boldsymbol{c}_j$, where $j \neq p$, so that cosine distance between them is maximum. We call this term "repulsive loss" similar to [27] we denote the loss as $\mathcal{L}_{\mathrm{rep}}$.

Pairwise term: term 4 in our objective maximizes the pairwise distance matrix of the centroid vectors. For the distance we proposed to use the squared euclidean distance and we denote the loss as $\mathcal{L}_{\mathrm{pair}}$.

Overall loss: is defined as

$$\mathcal{L} = \mathcal{L}^{\mathrm{cls}} + (\lambda_1\mathcal{L}^{\mathrm{attr}}) - (\lambda_2\mathcal{L}^{\mathrm{rep}}) - (\lambda_3\mathcal{L}^{\mathrm{pair}}) \tag{4}$$

where $(\lambda_1, \lambda_2, \lambda_3)$ are the coefficients. The hyper parameters $(\lambda_1, \lambda_2, \lambda_3)$ are important for minimizing the intra-class variation (to minimize the variance within the same class). More specifically, terms 3 and 4 focus on keeping the features of different classes separable, and term 2 focuses on minimizing the intra-class distances. All of them are essential to our model. We refer to the combination of the new added terms as the centroid loss, i.e., term 2, term 3, and term 4.

3.3 Explanations

We seek to identify a feature with a causal impact on the model prediction decision process. We follow [28]'s definition of intervention: an intervention is an idealized experimental manipulation carried out on some variable $\boldsymbol{x}$ which is hypothesized to be causally related to changes in some other variable p. Any intervention on the text input using attributions on the prediction p is a causal process that changes the model prediction. Therefore, if the intervention changes the model prediction, it is probably due to the adjustment in the causal space of the text input. We will use the idea of "intervention" to understand the effectiveness of our approach for both *why p?* and *why p and not q?*.

Why p? For this type of explanations, we identify potential causal attributes by calculating the cosine similarity between each $\boldsymbol{x}_i$ and the corresponding $\boldsymbol{c}_p$

of class p. A higher score indicates a more informative attribute. The negative scores indicate the features have negatively contributed to the specific class classification and vice-versa. For experiments, we intervene on the text input to remove irrelevant attributes, i.e., replacing each factor with a "<pad>" followed by observing the change in the model's probabilities.

Why p and not q?: Given any text instance, a classifier predicts p and a centroid vector $\boldsymbol{c}_p$. A p-contrast question is of the format 'Why [predicted-class (p)] not [desired class (q)]?'. By specifying the desired class, we limit our search space to a single alternative. Given the text input, we estimate attribution scores for "p" using $\boldsymbol{c}_p$. For the desired class q, we calculate the attribution scores of the text input using $\boldsymbol{c}_q$. Please note that, here we also use cosine similarity. We find the attribution scores for contrastive explanations as $\boldsymbol{z}_c = \boldsymbol{z}_p - \boldsymbol{z}_q$, where $\boldsymbol{z}_p$ is the attribution score for the predicted class p obtained using $\boldsymbol{c}_p$ and $\boldsymbol{z}_q$ is the attribution scores for the foil class q obtained using $\boldsymbol{c}_q$.We follow the intervention approach as in "*Why p?*," to find the candidate attributes for the contrastive explanation.

4 Experiments

To effectively evaluate our approach, we devise a measure to rank the identified causal attributes. Given a prediction "p," for "*why p*," we rank each attribute by how much it contributes to prediction "p" using $\boldsymbol{c}_p$. As for contrastive explanations, we rank each attribute using $\boldsymbol{z}$ by how contrastively useful it is to the model for choosing "p" against "q." All evaluations follow an interventionist approach defined in Section (3.3).

4.1 Setup

Datasets. We adopt the IMDB datasets [14] (train:25000, test:25000 samples) with binary labels, AG news [30](train:102080, test:25520 samples) with four classes, and YELP reviews [23] (train:110400, test:27600 samples) with binary labels. We hold out 10% of the training examples as the development set. We limit the length of the input to 50 for YELP and IMDB and 20 for AG news.

Model. The multi-head model [24] includes an embedding layer and multi-head attention layers. We tokenized sentences and randomly initialized the embedding layer and the centroid vectors. The dimension of the word embedding, centroid vector, and feature vector (at the output layer) is 128. For training the network, we use the Adam optimizer [10] with a batch size of 256 and a learning rate of 0.0001 (We have experimented with different values for the coefficient with interval 5 between 0 and 1000, for the experiments we used $(\lambda_1 : 1000, \lambda_2 : 10, \lambda_3 : 1000)$. The F1-scores for AG news topic classification, IMDB sentiment, and YELP review classification are summarized in Table 1. Performance is in terms of F1-score.

Table 1. Black-box(Multi-head) vs. intrinsic Multi-head neural network.

Models	Dataset		
	IMDB	YELP	AG news
Black-Box(Multi-head)	0.81	0.88	0.89
Proposed (Multi-head with centroid loss)	0.81	0.88	0.89

4.2 Explainability metrics

We adopt three metrics from prior work on evaluating word-level attribution (non-contrastive explanation): the area over the perturbation curve (AOPC) from ERASER [4], the log-odds scores [21, 3], and the degradation score to the trained model accuracy [17]. We also proposed new evaluation metrics for contrastive explanations. All the metrics measure the local fidelity by deleting or masking top-scored words.

4.3 Evaluating Why p?

We begin first by evaluating the faithfulness of " *Why p?*" questions. Faithfulness means the degree (trust of an explanation) to which an explanation influences the model prediction. ERASER proposes two metrics to measure the quality of the explanations:

Comprehensiveness: Measure whether all required features by the model to make a prediction are selected by the explanation method. To use this metric, we first need to compute a new sentence. For example, given an input text $\boldsymbol{X}$, the new sentence is defined as $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{R}$, where $\boldsymbol{R}$ is the set of salient features identified by the explanation method. Let $f_\theta(\boldsymbol{X})_p$ be the neural network output for class p. The measure of comprehensiveness is calculated as:

$$\text{Comprehensiveness} = f_\theta(\boldsymbol{X})_p - f_\theta(\tilde{\boldsymbol{X}})_p \tag{5}$$

A higher score implies that the identified tokens included in $\boldsymbol{R}$ were more influential in the model's predictions, compared with other tokens.

Sufficiency: The second metric focuses on evaluating whether the identified features were enough to predict the same label as using the full text or not, and is defined as follows:

$$\text{Sufficiency} = f_\theta(\boldsymbol{X})_p - f_\theta(\boldsymbol{R})_p \tag{6}$$

Under sufficiency metric, lower scores are better. We calculate the AOPC for both comprehensiveness and sufficiency using a variety of token percentages: $5\%, 10\%, 15\%, 20\%$, and 25%.

Log-odds: Log-odds score is calculated by averaging the difference of negative logarithmic probabilities on the predicted class over all of the test data before and after masking the top $m\%$ features with zero paddings,

$$\text{Log-odds}(m) = \frac{1}{t}\sum_{i=1}^{t} \log \frac{p(\hat{y}|\boldsymbol{X}_i^{(m)})}{p(\hat{y}|\boldsymbol{X}_i)}, \tag{7}$$

where $\boldsymbol{X}_i^{(m)}$ is the new input based on replacing the top $m\%$ with the special token `<pad>` in $\boldsymbol{X}_i$ and t is the total number of samples. Lower log-odd scores are better.**Degradation score**: Words are ranked according to "*why p?*" (defined in Subsection 3.3-*Why p?*). In this way, higher-ranked tokens (features) are recursively eliminated. The degradation score to the trained model accuracy is calculated. We perform this experiment using a variety of token percentages: $5\%, 10\%, 15\%, 20\%$, and 25%.

Results We compare our technique with competitive baselines, namely Shapely-based methods (L/C-Shapely) [3], using log-odds, AOPC, and degradation score. The log-odds and degradation scores are shown in Figure 3. The L/C-Shapley focuses on instance-wise feature importance scores. Shapley values are extremely expensive to compute and L/C-Shapley were proposed to compute approximate Shapley values. We evaluate the explanation on the test set of the datasets.

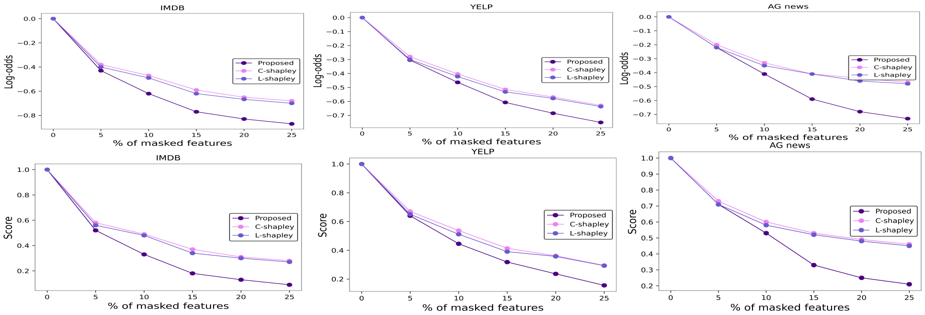

Fig. 3. Log-odds scores as a function of masked features (top). A steeper decline indicates a better performance.Degradation score (y-axis) as a function of removed tokens(bottom). A steeper decline indicates a better performance.

Our approach achieves the best performance on both metrics (log-odds and degradation score). Note that L-Shapley applying approximation Shapley values perform better than C-Shapley. The results also show that the neural network classifier employs less number of features for making predictions. Our method also outperforms Shapley approximation methods on ERASER metrics achieving the best result (Table 2) for both comprehensiveness and sufficiency on the three datasets.

4.4 Evaluating Why p and not q?

To evaluate the faithfulness of contrastive explanations, we use the following metrics:

Contrastive overlap score (COS) (%): we calculate the overlap (%) between the sets of causal attributes of "*Why p?*" and "*Why p and not q?*. Lower % indicates more difference between the explanations of "*Why p?*" and "*Why p and not q?*.

Table 2. Eraser benchmark scores: Comprehensiveness and sufficiency in terms of AOPC

	L-Shapley	C-Shapley	Proposed
IMDB			
Comprehensiveness	0.575	0.554	0.704
Sufficiency	0.1722	0.172	0.112
YELP			
Comprehensiveness	0.494	0.479	0.562
Sufficiency	0.172	0.172	0.112
AG news			
Comprehensiveness	0.384	0.37	0.524
Sufficiency	0.247	0.246	0.086

Contrastive confidence score (CCS): For a confidence score, we analyze the change in the probability of the contrastive class "*q*." We remove the attributes that distinguish "*p*" from "*q*" in order of their importance, until the model's prediction is flipped to another class. Please note the scores of the features are obtained using "*why p and not q?*" We calculate the difference in the probability of"*q*" before and after the intervention. An increase in the probability indicates an informative contrastive explanation.

Contrastive gain (CAG): This metric measures the quality of contrastive explanations compared to the non-contrastive explanations. Here, our explanations for the question "*why p?*" will be called non-contrastive explanations. Given a prediction "*p*" and foil "q," we measure the change in the probability score of "*q*" after removing salient features using attribution-scores obtained from "*why p?*" and also from "*why p and not q?*" explanation. We use our approach as the baseline for "*why p?*", because our method outperformed [3].For the "*why p and not q?*" explanation, we used the method described in Section (3). A higher contrastive gain indicates that our contrastive explanation is better in answering "*why p and not q?*" questions. In summary, the contrastive gain measures the change in probability of the foil class after removing some features.

Results We use the AG news dataset to evaluate our contrastive explanation method. For contrastive overlap (COS), the results in Figure 4 show that most contrastive explanations do have fine-grained differences from "*Why p?*" questions. The result suggests that the model is not using the same reasoning for "*Why p?*" when answering the contrastive questions. We observed that, for multi-class problems, there are fine-grained differences between "*Why p?*" and "*Why p and not q?*" compared to a binary problem such as sentiment classification where there might be a higher similarity between the two explanations.

For (CCS), results shown in Table 3 indicate the effectiveness of our approach in finding contrastive information. Meaning that there is an increase in the score of the foil "*q*" when removing the features that distinguish "*p*" from "*q*". We also show the scores of other classes when using the (CCS) metric. We re-trained the same model again on AG news and re-calculated the CCS. The results are shown in Table 4. We can see that when the foil *q* is set to "business"

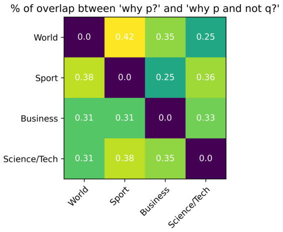

Fig. 4. Overlap score between "*why p?*" and "*why p and not q?*" questions. "0.0" means that we did not consider the contrastive explanations when "p" and "q" are the same (X-axis: refers to why p? questions and Y-axis: refers to why p and not q? questions.)

Table 3. Constrative confidence score (CCS). Empty cells mean that we cannot find a contrastive explanation for the same class i.e., the foil should be different from the predicted class. The highlighted cells show the scores of the foil after removing the salient features.

	World(q)		Sport(q)		Business(q)		Sci/Tech(q)	
Class(p)	Before	After	Before	After	Before	After	Before	After
World			0.05	0.22	0.05	0.41	0.01	0.33
Sport	0.07	0.45			0.01	0.35	0.04	0.21
Business	0.06	0.38	0.003	0.16			0.13	0.37
Sci/Tech	0.02	0.32	0.04	0.12	0.13	0.52		

Table 4. We compare the scores of other classes when evaluating the CCS for the foil. Here the foil is the business class.

	World	Sport	Business(q)	Sci/Tech
World(p)	0.2	-0.8	0.4	0.1
Sport(p)	0.3	0.1	0.5	-0.8
Sci/Tech(p)	-0.7	0.1	0.4	0.1

and evaluated with different classes (p) such as "world, sport, sci/tech ."The probability score for the "business" is higher compared to other classes when using *why p and not business.?* Due to page limit, we only show the results for the "business" class (see Table 4). Figure 5 compares our non-contrastive explanations and contrastive explanation methods (CAG). We use the AG news data and plot the results for different "*why p and not q?*" questions. The results in Figure 5 indicate that our contrastive explanations are better capturing the features that contribute prediction of 'not q' than non-contrastive explanations, especially when there are more fine-grained differences. The results show that non-contrastive explanation is not always achieving high contrastive scores when top features are masked. Instead of tracking the change in probability socre of

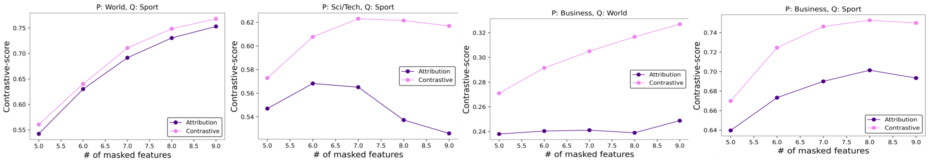

Fig. 5. Conrastive gain as a function of removed tokens. A higher gain indicates that the method was better in capturing contrastive information. Attribution refers to our non-contrastive method.

"q" after removing salient as in contrative gain , we instead calculate the AOPC using different percentages $(25\%, 30\%, 35\%, 40\%, 45\%)$. The results are summarized in Table 5. Our contrastive-explanation has the highest AOPC compared to our non-contrastive explanation method.

Table 5. Contrastive gain (CAG): Evaluating the effectiveness of using contrastive explanation when there are fined grained differences. We use different percentages $(25\%, 30\%, 35\%, 40\%, 45\%)$ to calculate the AOPC.

P	**Q**	AOPC(non-contrastive)	AOPC(contrastive)
World	Business	0.04	0.065
Business	Sci/tech	-0.001	0.002
Sci/tech	World	0.304	0.341
Sci/tech	Business	0.056	0.058
Sport	Business	0.05	0.052
Sport	Sci/tech	0.006	0.009

Highlighting *why p and not q?* questions: We show qualitative results for interpreting the model predictions using our proposed approach; for example, answers to the "*Why p?*" and "*Why p and not q?*" questions are shown in Table 6. These results show that the model implicitly learns the contrastive information when making the prediction.

Contrastive explanations applied to sentiment classification. For a contrastive explanation, if there are no fine-grained differences between "p" and "q", then the same reasoning used for "*why p?*" questions will also be used to answer "*why p and not q?*" questions. We observed this behavior in binary text classification. For instance, we found that the model uses the same reasoning for both questions (see Table 7). We attribute this observation to the fact that "*why p and not q?*" cites the causal difference between p and *not*-q, i.e., consisting of a cause of p and the absence of a corresponding event in the history of q. We also found that explaining "*Why p and not q?*" is not the same as explaining "*Why q and not p?*." In the case of sentiment classification, we found that these two questions provide different answers, and it is consistent with the work of [12]. To validate our observations in sentiment classification, we focus now on the

Table 6. Contrastive explanations on AG news.

Text	World	Sport	Business	Sci/tech
record shown mutilated body found iraq kidnapped aid worker margaret hassan british official say still believe british irish citizen dead(P:World,Q:others)	iraq	dead	hassan	kidnapped
search war begin today software giant microsoft unveils test version new search engine looking remarkably like one chief rival google. (P:Sci/tech,Q:others)	engine	microsoft	engine	version
version desktop search tool computer run apple computer mac operating system google chief executive eric schmidt said friday(P:Sci/tech,Q:others)	desktop	apple	schmidt	mac
inflation dozen nation sharing euro slowed initially estimated september company reduced price lure customer store offsetting record energy cost.(P:World,Q:others)	store	inflation	price	lure

overlap between "*why p and not q?*" and "*why q and not p?*." We use the IMDB dataset and calculate the overlap between the attributes (minimum subset of the attributes required to flip the prediction) of "*why p and not q?*" and "*why q and not p?*." The ratio of similarly was zero, which means the explanations are entirely different.

4.5 Deep learned features

In Figure 6, we apply PCA over the sentence vectors learned via our proposed method. The centroid loss forces the network to learn meaningful representations for the embedding layer. We can see that our current model struggles with negative sentiment reviews according to the number of points (yellow color) appearing in the positive sentiment cluster.

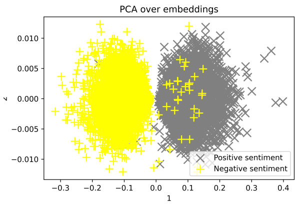

Fig. 6. The distribution of deeply learned features under the centroid loss on IMDB dataset.

Table 7. *why p and not q?* Contrastive explanation is the same as *why p?* explanations in binary sentiment classification.

Text	Highlight
the story is enjoyable and easy to follow this could have been easily messed up but the actors and director do a great job of keeping it together the actors themselves are fantastic displaying wonderful character and doing a terrific job gotta find a copy somewhere (P:positive,Q:negative)	fantastic
this performance that should elevate the film to a platform where it a place on the best ever lists of courtroom dramas however despite its apparent obscurity sergeant still remains a taut and compelling examination like a book that you just can't put down highly recommended (P:positive,Q:negative)	recommended
imagined in my mind what i saw on screen was slightly different however it wasn't enough to make me dislike the mini series i recommend this for anyone who has read the novel you will not be disappointed if you have 8 out of 10 stars (P:positive,Q:negative)	8
provide someone to at well one must do something beside during this film the movie is being sold on vhs now by people on e bay spare yourself the expense and the waste of time a comedy without a laugh a musical without a memorable song or dance (P:negative,Q:positive)	waste

4.6 Discussion

Intrinsic models We have introduced an approach for constructing interpretable neural models. We have shown that introducing additional constraints to the learning objective does not sacrifice performance, and it also provides faithful explanations to the black-box predictions. The centroid vectors are used as a proxy to explain the predictions. We found that discriminative features (words) tend to get closer to the corresponding centroid vector and irrelevant features tend to get further away. Discriminative features are the words employed by the network to make a prediction, and irrelevant features are the tokens ignored by the classifier when making a prediction.

Centroid loss The empirical results demonstrated the usefulness of the centroid vectors in finding the most salient features for every input. The centroid loss does not require complex recombination of the training samples. Our approach targets the learning objective of the intra-class using term 2, which is very beneficial to discriminative feature learning. We have also shown that our contrastive explanations are helpful when there are fine-grained differences.

5 Conclusion

We have proposed an intrinsic neural model capable of explaining its predictions faithfully. Our network architecture relies on a centroid loss to learn centroid vectors. These centroid vectors are then used to provide two types of explanations: (i) non-contrastive explanations and (ii) contrastive explanations. Our feature attribution method provides a better faithful explanation than Shapley's approximation based on three datasets using three metrics. We have also proposed additional metrics to evaluate contrastive explanations. Our contrastive explanation method can provide additional insights to non-contrastive explanation, resulting in a better understanding of the neural model predictions. We have also shown that interpretability does not affect the predictive accuracy of the neural network. In future work, we would like to study the use of our intrinsic neural model with different tasks in the NLP domain and extend our current solution for producing counterfactual explanations.

Acknowledgements

We would like to acknowledge the support of the Alberta Machine Intelligence Institute (Amii), and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Bashier, H.K., Kim, M.Y., Goebel, R.: Disk-csv: Distilling interpretable semantic knowledge with a class semantic vector. In: Proceedings of the 16th Conference of the EACL Main Volume. pp. 3021–3030 (2021)
2. Bastings, J., Aziz, W., Titov, I.: Interpretable neural predictions with differentiable binary variables. In: Proceedings of ACL. pp. 2963–2977 (2019)
3. Chen, J., Song, L., Wainwright, M.J., Jordan, M.I.: L-shapley and c-shapley: Efficient model interpretation for structured data. ICLR 2019 (2018)
4. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: Eraser: A benchmark to evaluate rationalized nlp models. In: Proceedings of the 58th ACL. pp. 4443–4458 (2020)
5. Einhorn, H.J., Hogarth, R.M.: Judging probable cause. Psychological Bulletin 99(1), 3 (1986)
6. Hendricks, L.A., Hu, R., Darrell, T., Akata, Z.: Generating counterfactual explanations with natural language. In ICML Workshop on Human Interpretability in Machine Learning, pages 95–98 (2018)
7. Hilton, D.J.: Conversational processes and causal explanation. Psychological Bulletin 107(1), 65 (1990)
8. Ismail, A.A., Corrada Bravo, H., Feizi, S.: Improving deep learning interpretability by saliency guided training. Advances in Neural Information Processing Sys- tems 34, 26726–26739 (2021)
9. Jacovi, A., Swayamdipta, S., Ravfogel, S., Elazar, Y., Choi, Y., Goldberg, Y.: Contrastive explanations for model interpretability. arXiv preprint arXiv:2103.01378 (2021)
10. Kingma, D.P., Ba, J.L.: Adam: A method for stochastic optimization. cornell university library. arXiv preprint arXiv:1412.6980 (2017)
11. Koura, A.: An approach to why-questions. Synthese 74(2), 191–206 (1988)
12. Lipton, P.: Contrastive explanation. Royal Institute of Philosophy Supplements 27, 247–266 (1990)
13. Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3), 31–57 (2018)

14. Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y., Potts, C.: Learning word vectors for sentiment analysis. In: Proceedings of ACL. pp. 142–150. ACL (2011)
15. McGill, A.L., Klein, J.G.: Contrastive and counterfactual reasoning in causal judgment. Journal of Personality and Social Psychology 64(6), 897 (1993)
16. Miller, T.: Explanation in artificial intelligence: Insights from the social sciences. Artificial intelligence 267, 1–38 (2019)
17. Nguyen, D.: Comparing automatic and human evaluation of local explanations for text classification. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1069–1078 (2018)
18. Rathi, S.: Generating counterfactual and contrastive explanations using shap. arXiv preprint arXiv:1906.09293 (2019)
19. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1135–1144. ACM (2016)
20. Rudin, C.: Please stop explaining black box models for high stakes decisions. 32nd Conference on Neural Information Processing Systems (NIPS 2018), Workshop on Critiquing and Correct- ing Trends in Machine Learning. (2018)
21. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. pp. 3145–3153. JMLR. org (2017)
22. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Proceedings of International Conference on Machine Learning (ICML). p. 3319–3328 (2017)
23. Tang, D., Qin, B., Liu, T.: Document modeling with gated recurrent neural network for sentiment classification. In: Proceedings of the 2015 conference on EMNLP. pp. 1422–1432 (2015)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
25. Waa, J.v.d., Robeer, M., Diggelen, J.v., Brinkhuis, M., Neerincx, M.: Contrastive explanations with local foil trees. arXiv preprint arXiv:1806.07470 (2018)
26. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv. JL and Tech. 31, 841 (2017)
27. Wang, Y., Huang, H., Rudin, C., Shaposhnik, Y. Understanding How Dimension Reduction Tools Work: An Empirical Approach to Deciphering t-SNE, UMAP, TriMap, and PaCMAP for Data Visualization. J. Mach. Learn. Res., 22(201), 1-73 (2021)
28. Woodward, J.: Making things happen: A theory of causal explanation. Oxford university press (2005)
29. Yang, L., Kenny, E., Ng, T.L.J., Yang, Y., Smyth, B., Dong, R.: Generating plausible counterfactual explanations for deep transformers in financial text classification. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6150–6160 (2020)
30. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. In: Advances in Neural Information Processing Systems. pp. 649–657 (2015)